IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2020, accepted January 28, 2020, date of publication February 3, 2020, date of current version February 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2971004

Restricted Region Based Iterative Gradient Method for Non-Targeted Attack

ZHAOQUAN GU[1], WEIXIONG HU[1], CHUANJING ZHANG[1], LE WANG[1], CHUNSHENG ZHU[2,3], AND ZHIHONG TIAN[1]

[1]Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510006, China
[2]SUSTech Institute of Future Networks, Southern University of Science and Technology, Shenzhen 518055, China
[3]PCL Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen 518055, China

Corresponding author: Le Wang (wangle@gzhu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61902082, Grant U1636215, and Grant 61572492, in part by the project "PCL Future Regional Network Facilities for Large-scale Experiments and Applications" under Grant PCL2018KP001, and in part by the Guangdong Province Universities and Colleges Pearl River Scholar Funded Scheme (2019).

ABSTRACT Neural networks have been widely applied but they are still vulnerable to adversarial examples. More and more defense models have been proposed and they can resist the attacks to the neural networks. In order to generate adversarial examples with good transferability, we propose the restricted region based iterative gradient method (RRI-GM) for non-targeted attack, which aims at generating adversarial examples to make black-box defense models output wrong decision. We first use object detection algorithm to restrict some key regions in the images, since we regard perturbation in the key region affects more than the whole image. To improve the efficiency of attacks, we use gradient-based attack methods and they show good performance. In addition, the process is iterated for multiple rounds to generate adversarial examples with good transferability. Furthermore, we conduct extensive experiments to validate the effectiveness of the proposed method, and the results show that our method can achieve good attack performance against black-box defense models.

INDEX TERMS Adversarial examples, black-box attack, transferability, restrict region, gradient-based attack, non-targeted attack.

I. INTRODUCTION

With the fast development of deep learning, neural networks have achieved great success in a large number of applications [1], [2]. However, deep neural networks (DNNs) are highly vulnerable to *adversarial examples* [3]–[6]. These maliciously generated adversarial examples add small perturbations to the original images, which cannot fool people but make DNNs output incorrect or unreasonable predictions. Adversarial examples may also exist in real world [7]–[9], which have caused a wide range of security concerns in many sensitive applications, such as self-driving cars, finance evaluation and FacePay [10]–[12].

Adversarial examples have attracted much attention in recent years because they can serve as an important surrogate to evaluate the robustness of neural networks [13], [14]. Fast gradient sign method (FGSM) is the pioneering work that generates adversarial examples to fool the DNNs [3]. After that, many gradient-based methods are proposed to generate adversarial examples according to the varied gradients. Some typical works include basic iterative method [6], project gradient descent method [13], Carlini & Wagner's method [15] and momentum iterative method [16]. These methods need to know the gradient information of the specific neural network they attack, which are referred to as *white-box* attacks. Without the information of the neural networks, some methods can also attack the networks with high success rate, which are referred to as *black-box* attacks. A common strategy of black-box attack is to utilize the cross-model transferability of adversarial example [17], [18], which implies the generated adversarial examples that fool a white-box neural network can also fool a black-box neural network with high probability. The transferability enables practical *black-box* attacks to real-world applications without acquiring the neural networks' information; hence the adversarial examples might induce serious security problems in practice.

The associate editor coordinating the review of this manuscript and approving it for publication was Kim-Kwang Raymond Choo.

Realizing the existence and effects of adversarial examples, more and more researchers turn to build robust neural networks that can defend the adversarial examples. For example, many works introduce *adversarial training* as an effective defense method [19], [20], which utilizes the generated adversarial examples to train neural networks. These works are shown to achieve good defense results against white-box attacks. *Ensemble learning* is another strategy which combines multiple neural networks to defend the adversarial example [21]. This method could achieve good results under some circumstances, but it is highly related to the integrated neural networks. There are also some other methods such as network distillation [22] and input reconstruction [23], but they cannot defense against all kinds of adversarial examples.

In this paper, we explore efficient black-box attack methods against DNNs. Utilizing the transferability of adversarial examples, we are able to attack black-box neural networks, but the success rate is very low especially against the ensemble defense method. In order to generate robust adversarial examples that evade both normally trained neural networks (white-box) and defense neural networks (black-box), we propose the restricted region based iterative gradient method. The adversarial attacks are classified into *non-targeted attack* and *targeted attack* according to different goals of the attacks. Non-targeted attacks tend to make the neural networks make wrong classification of the adversarial example, while targeted attacks aim at misclassifying the adversarial example as a specific target. In this paper, we introduce our method for non-targeted attack; with small modification, this method can be also applied to targeted attack.

There are three insights in designing the restricted region based iterative gradient method. In the first place, most methods modify the whole image to generate adversarial examples, however we pay more attention to the discriminative regions in images and we only alter some key restricted regions other than the whole image. Second, when we generate adversarial examples by adding perturbation to each pixel, it is more important to choose the right perturbation direction other than compute the specific perturbation value. This intuition is also adopted in FGSM attack [3]. Finally, neural networks would utilize more training iterations to achieve good performance, adversarial examples should also be generated in an iterative way to attack the neural networks with high success rate. Combining these aspects, we add perturbation to the restricted region iteratively to generate adversarial examples against black-box neural networks. We have conducted extensive experiments on the CAAD (competition on adversarial attacks and defenses) dataset (the selected images are from ImageNet dataset). The results show that the proposed restricted region based iterative gradient attack method helps to improve the success rate of black-box attack against the normally trained models and defense models by a large margin.

The remainder of the paper is organized as follows. The next section highlights the related works in adversarial attacks and defense methods. We introduce the preliminaries in Section III. We present the methodology in Section IV and describe the experimental results in Section V. The advantages and disadvantages of the proposed method are discussed in Section VI. Finally, we conclude the paper in Section VII.

II. RELATED WORK

In this section, we introduce the related works in generating adversarial examples against DNNs and some defense methods against such attacks. DNNs have beed widely applied in many fields such as image classification [24]–[26], text processing [27] and speech recognition [28]. Although DNNs could achieve good performance, they have some intrinsic problems that could cause serious security concerns. Many works focus on attack image classification models which add small perturbation to the original images and these generated images could cause the DNNs make wrong prediction. These methods are called *image domain attacks*. A few works explore whether such adversarial images exist in the physical world and these methods are called *physical domain attacks*. In this section, we present some representative attack methods and introduce some defense intuitions against the attacks.

A. ADVERSARIAL ATTACK: IMAGE DOMAIN

Adversarial attacks in the image domain imply the attacks are conducted to the original images and the generated adversarial examples could fool image processing neural networks (such as image classification and object detection). There are two types of the attack methods: white-box attacks assume the neural networks' information is known beforehand, while black-box attacks do not need to know these information.

1) WHITE-BOX ATTACK

White-box attacks assume the full knowledge of the neural networks is known beforehand, including the structure of the network model, all the trained parameters, etc. Many white-box attack methods adjust the gradients to generate adversarial examples [3], [5], [6], [13], [15], [16], [29]–[34]. Fast Gradient Sign Method (FGSM) [3] is the pioneering work that generates adversarial examples by adding gradient noise to the original images with only one step, but the attack is less effective. In [13], an iterative FGSM algorithm called project gradient descent (PGD) is proposed, which is considered as one strongest first-order iterative gradient attack method. In the 2017 NIPS adversarial examples competitions, momentum iterative fast gradient sign method (MI-FGSM) [16] modifies the FGSM algorithm with momentum, which achieves good results in the competition. The aforementioned algorithms can be formally stated as white-box attack methods since they utilize the gradients during training. Besides the gradient-based algorithms, some optimization based attacks are also proposed. For example, Carlini & Wagner (C&W) attack [15] is proposed to adjust the generated disturbance with optimization method. DeepFool is proposed in [29], which iteratively calculates the closest boundary to the original image and then generates the

adversarial examples. These attack methods are summarized in Table 2.

2) BLACK-BOX ATTACK

Black-box attacks have no access to the neural networks' information, such as the model parameters or the gradients [17], [35]–[39]. In [4], it is shown that the adversarial examples can be generalized between different different DNNs. In other words, if the adversarial example can fool some neural network, it can also fool another neural network with high probability. This good transferability of the adversarial examples is normally utilized to attack black-box neural networks. In [35], the methods are proposed to improve the transferability, which enable powerful black-box attacks. In [17], it uses queries to distill the knowledge of the black-box model and train a surrogate model; then it turns black-box attack to white-box attack.

B. ADVERSARIAL ATTACK: PHYSICAL DOMAIN

Adversarial attacks in the image domain cannot be realized in the real environment. This is because many attacks methods would add perturbation to the whole image but the generated adversarial examples cannot be fulfilled in practical applications. Therefore, some works explore whether the adversarial examples could occur in the real world. In [6], it verifies the existence of real adversarial examples. By printing the adversarial images generated in the image domain on the paper, it uses a mobile phone camera to capture the images and the neural network also misidentifies a washing machine as a speaker. In self-driving filed, robust physical perturbation (RP2) method is proposed in [8], which generates black and white stripes that are pasted on the road sign. This method could cause the neural network recognize the stop road sign as speed limit. After that, the RP2 method is extended in [40], which pastes colored stripes on the stop road sign and can cause the object detection neural networks fail to detect the road sign. In face recognitions, the method proposed in [41] can design special glasses to attack the face recognition system. When the attacker wears the glasses, two mainstream face recognition neural networks would make wrong recognition. There are some other methods [42], [43] and we also list them in Table 2.

C. ADVERSARIAL DEFENSE METHODS

Many methods have been proposed to increase the robustness of DNNs against the adversarial attacks. Adversarial training is a common adopted strategy that can defend against adversarial attacks. The intuitive idea is to train the neural networks with generated adversarial examples and they could show good performance against the corresponding attacks. Besides adversarial training, there are some other proposed methods that could improve the robustness of the neural networks. In [20], it proposes obfuscated gradients method to improve the robustness. In [22], distill defense method is also introduced against the adversarial examples.

III. PRELIMINARIES

In this section, we introduce the notations of adversarial attack and formulate the problem formally. Since our proposed method is highly related to gradient based attack methods, we also introduce some typical gradient based attacks.

A. SYSTEM MODEL AND PROBLEM DEFINITION

Considering the image classification task where a set of images are denoted as $X = \{x_1, x_2, \ldots, x_n\}$ and their corresponding classification labels are denoted as $Y = \{y_1, y_2, \ldots, y_n\}$. Suppose a neural network M is trained on the set of images and their labels; it can predict a new image x' with a correct classification label y' with high probability. For simplicity, we denote $f_M(\cdot)$ as the classifier output of the neural network, which implies $f_M(x') = y'$.

Adversarial attacks generate new images that fool the trained neural network. Let x_{true} denote a true (clean) example and y denote the corresponding ground-truth label, i.e. $f_M(x) = y$. The goal is to generate an adversarial example x_{adv} to fool the classifier with small perturbation, i.e. $f_M(x_{adv}) \neq y$. There are many methods to evaluate the perturbation and L_p norm [3]–[6] is commonly utilized to restrict the perturbation within a small range:

$$||x_{adv} - x_{true}||_p \leq \epsilon,$$

where ϵ is a pre-defined threshold. In this paper, we also use the L_p norm to evaluate the perturbation.

Notice that, if the information about the trained neural network M is known beforehand, we call this white-box attack. Otherwise, it is called black-box attack without such kind of information. Considering white-box attacks, denote $J(x_{adv}, y)$ as the loss function that image x_{adv} is classified as the right label y; the goal is to optimize

$$arg\max_{x_{adv}} J(x_{adv}, y), \quad s.t.\ ||x_{adv} - x_{true}||_p \leq \epsilon. \tag{1}$$

Many works utilize the transferability to attack black-box neural networks; in this paper we also generate adversarial examples on the basis of white-box networks and adopt them to attack black-box networks.

In addition, if the goal is to make $f_M(x_{adv}) \neq y$, it is called non-targeted attack; on the contrary, targeted attack intends to satisfy $f_M(x_{adv}) = y_{adv}$ where $y_{adv} \neq y$ is a pre-defined target. For targeted attack, we can modify the loss function as $J(x_{adv}, y_{adv})$ and minimize the function under the perturbation constraint. In this paper, we focus on non-targeted attack and the method can be easily extended to targeted attack.

B. GRADIENT-BASED ADVERSARIAL ATTACK METHODS

There are some typical gradient-based adversarial attack methods solving the optimization problem in Eqn. (1).

Fast Gradient Sign Method (FGSM) is a pioneering work [3], which generates an adversarial example x_{adv} by linearizing the loss function in the input space and performing one-step update as

$$x_{adv} = x_{true} + \epsilon \cdot sign(\nabla_x J(x_{true}, y)). \tag{2}$$

TABLE 1. Attack methods against neural networks.

Attack Method	Domain	White or Black Box	Description
FGSM [3]	Image domain	White-box	Attack Recognition
L-BFGS [4]	Image domain	White-box	Attack Recognition
BIM [5]	Image domain	White-box	Attack Recognition
ILCM [6]	Image domain	White-box	Attack Recognition
PGD [13]	Image domain	White-box	Attack Recognition
C&W [15]	Image domain	White-box	Attack Recognition
MI-FGSM [16]	Image domain	White-box	Attack Recognition
DeepFool [29]	Image domain	White-box	Attack Recognition
JSMA [30]	Image domain	White-box	Attack Recognition
ATN [31]	Image domain	White-box	Attack Recognition
DAG [32]	Image domain	White-box	Attack Detection
Universal Attack [33]	Image domain	White-box	Universal Adversarial Attack
Feature Adversary [34]	Image domain	White-box	Use internal neural network layers
One Pixel Attack [36]	Image domain	Black-box	Attack Recognition
UPSET [37]	Image domain	Black-box	Attack Recognition
ZOO [38]	Image domain	Black-box	Attack Recognition
LocalSearchAttack [39]	Image domain	Black-box	Attack Recognition
RP2 [8]	Physical domain	White-box	Attack Traffic Sign
Extended RP2 [40]	Physical domain	White-box	Attack Traffic Sign
GAN [42]	Physical domain	White-box	Attack Face Recognition
AGNs [42]	Physical domain	White-box	Attack Face Recognition
ADVHAT [43]	Physical domain	White-box	Attack Face Recognition

In Eqn. (2), ϵ is a hyperparameter that can be used to control the perturbation. $sign(\cdot)$ is the sign function and the perturbation meets the L_∞ norm constraint. $\nabla_x J(\cdot,\cdot)$ is the gradient of the loss function and the method generates the adversarial examples according to the gradient.

Basic Iterative Method (BIM) extends FGSM by iteratively applying gradient updates multiple times with a small step size α, which can be expressed as

$$x_{adv}^{t+1} = x_{adv}^{t} + \alpha \cdot sign(\nabla_x J(x_{adv}^{t}, y)),$$

where $x_{adv}^{0} = x_{true}$. To restrict the generated adversarial examples within the ϵ-ball of x_{true}, the method can clip x_{adv}^{t} after each update, or set $\alpha = \epsilon/T$ where T is the number of iterations. It has been shown that BIM induces much more powerful white-box attacks than FGSM at the cost of worse transferability.

Project Gradient Descent (PGD) is also an iterative attack method [13]. It can be regarded as the advanced fast gradient sign method. The PGD attack consists of the following steps: it first initializes the search for an adversarial example at a random point within the allowed norm ball, then it runs several iterations of the basic iterative method to find the adversarial example. The process can be formulated as

$$x^{t+1} = \prod_{x+S}(x^{t} + \alpha sgn(\nabla_x L(\theta, x, y))).$$

The generated noisy initial point could help conduct stronger attack than other previous iterative methods such as BIM.

Carlini & Wagner's (C&W) method is a powerful optimization-based method [15], which solves

$$arg\min_{x_{adv}} ||x_{adv} - x_{true}||_p - c \cdot J(x_{adv}, y). \tag{3}$$

The loss function $J(\cdot,\cdot)$ could be different from the cross-entropy loss and it can generate adversarial examples with smallest perturbation.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) can improve the transferability of adversarial examples [16]. This method assumes that perturbation in every epoch is related not only to the current gradient, but also to the previous calculated gradient. Then the update procedure can be formulated as

$$g^{t+1} = \mu \cdot g^{t} + \frac{\nabla_x J(x_{adv}^{t}, y)}{||\nabla_x J(x_{adv}^{t}, y)||_1},$$
$$x_{adv}^{t+1} = x_{adv}^{t} + \alpha \cdot sign(g^{t+1}),$$

where g^t gathers the gradient information up to the t-th iteration with a decay factor μ.

IV. METHODOLOGY

In the section, we describe the proposed method. Although there are many attack methods generating adversarial examples that can be utilized to attack black-box neural networks. The attack success rate of those adversarial examples is low, especially against the defense (black-box) networks that utilize the generated adversarial examples for training. In order to improve the transferability of adversarial examples, we propose restrict region based iterative gradient attack method (RRI-GM for short).

In the first place, in order to reduce perturbation and improve attack success rate, we only modify the key region in the image. Then we use gradient-based attack method and focus on alter perturbation direction to improve the efficiency of attack. In order to generate adversarial examples that

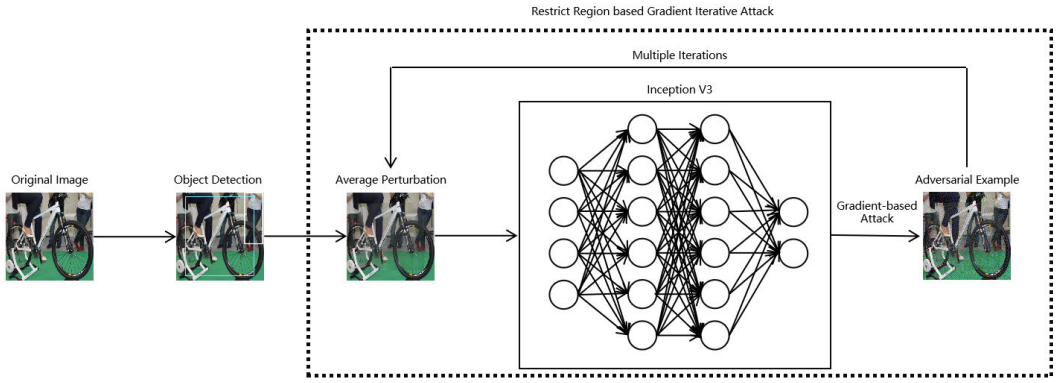


FIGURE 1. The process of the restrict region based iterative gradient attack method (RRI-GM).

FIGURE 2. Examples from IJCAI-AAAC2019 dataset for object detection by YOLO V3 algorithm.

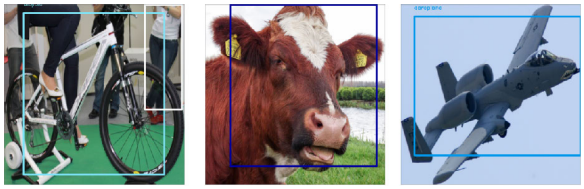


FIGURE 3. Object detection using YOLO V3 method; the examples are from CAAD competition.

have transferability to attack black-box defense networks, we iteratively generate the adversarial examples. As shown in Fig. 1, we use object detection algorithm to detect key region as the restricted region. Then, we use some perturbation methods in the key region, such as Guassian noise, average perturbation, etc. Table 3 shows the result of prediction accuracy of these perturbation methods (please refer to Section V). Through the comparison in our experiments, we find out that the average perturbation method outperforms other methods by about 3% $\sim$ 8%. Hence, in our proposed RRI-GM method, we select average perturbation method in the detected restricted region. After that, we regard these images (output by the average perturbation) as the input of InceptionV3 [26] for adopting FGSM to generate adversarial examples. Finally, we repeat the above process iteratively to generate adversarial examples that have good transferability and attack performance.

A. RESTRICT REGION

We use the object detection algorithm to find the key region of a image. We can only add perturbation to the key region of a image and this is the reason we call restricted region in our method. In Fig. 2, we show some examples of detected object by traditional object detection algorithm. We use YOLO-V3 as the detection algorithm and the examples are from the dataset of IJCAI-AAAC2019 competition.[1] We also adopt the detection algorithm on some images of CAAD competition[2] and these images are from ImageNet dataset. These figures show that the restricted regions can be generated and they represent the important parts in the images.

From these figures, we find out that the detected regions are not very neat and some images can even circle multiple boxes. In our algorithm, we select most selected box as the key region. We use four methods to conduct the restricted region attack and we restrict the maximum modification value as 16 to reduce the perturbation for people's vision, while implies the maximum changed value of a pixel cannot exceed 16. The four attack methods are described as follows:

- Guassian noise: we use a random Guassian distribution as the perturbation;
- Maximum value addition (Max-Value-Add for short): we add 16 (the maximum perturbation value) to every pixel value in the restrict region;
- Maximum value substraction (Max-Value-Sub for short): we subtract 16 (the maximum perturbation value) to every pixel value in the restrict region;
- Average perturbation (Avg-Perturbation for short): we modified the pixel value in the restrict region by calculating an average value of the surrounding pixels.

We show some examples in Fig. 4 after conducting different perturbation attacks in the detected region on ImageNet dataset. This first column shows the original images in the dataset; the next column shows the generated images by the Guassian noise attack; the third column shows the generated images by the Max-Value-Add method; the fourth column shows the generated images by the Max-Value-Sub method; and the last column shows the images by the average perturbation method. As we can see from these images, we can recognize them easily while the neural networks may make wrong prediction.

[1] https://tianchi.aliyun.com/competition/entrance/231701/introduction
[2] http://www.geekpwn.org/en/index.html

TABLE 2. Image classification results of ten models.

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
Accuracy	99.866%	93.166%	92.533%	92.633%	92.100%
Defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
Accuracy	**99.966%**	92.133%	92.266%	91.900%	93.333%

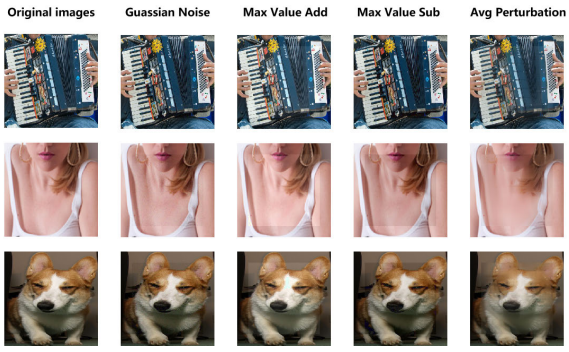


FIGURE 4. Original images and adversarial examples after adding different perturbation attacks (Guassian noise, maximum value addition, maximum value substraction, and average perturbation) in the detected region.

B. GRADIENT-BASED ATTACK

In our algorithm, we need to design or select a gradient-based attack method to improve our efficiency of attack. Our method is very scalable and we can use any gradient-based method if time permits. In this paper, we use gradient-based methods including FGSM, PGD, and MI-FGSM. Moreover, in order to improve transferability of generated adversarial examples, we use multiple iterations to conduct experiments and the generated examples show good transferability in attacking black-box models. As depicted in Fig. 1, each iteration consists of two main steps: the first step adopts the restricted region perturbation method while the second step utilizes gradient based methods on a trained network (Inception V3) to generate adversarial examples. In our experiments, we find out that the generated adversarial examples with more iterations could achieve better attack performance, while it costs more time. We will show these results in the following section.

V. EXPERIMENTS

We select 3000 images from the ImageNet dataset to conduct experiments and this dataset is used in the CAAD 2019 CTF image adversarial competition. We use 5 normally trained models (neural networks without adversarial learning) and 5 defense models (coupling with adversarial learning) which are shown to be robust against black-box attacks on the dataset. These normally trained models are:

- InceptionV3: it improves the network structure of Inception Module and introduces the idea of Factorization into small convolutions [44];
- InceptionV4: it combines Inception and ResNet [44];
- ResNetV2-152, ResNetV2-101, ResNetV2-50: ResNet is proposed in [45] and batch normalization is adopted in each layer for ResNetV2. The three networks represent 152, 101, 50 layers respectively.

We also choose 5 defense models as follows:

- AdvInceptionV3: it uses the adversarial examples against InceptionV3 model for adversarial training;
- Ens3AdvInceptionV3, Ens4AdvInceptionV3: they ensemble 3 or 4 models for adversarial training;
- AdvInceptionResNetV2: it uses adversarial examples against InceptionResNetV2 model for training;
- EnsAdvInceptionResNetV2: it ensembles 3 models against InceptionResNetV2 for adversarial training.

In our experiments, we choose different gradient-based methods, including fast gradient sign method (FGSM), project gradient descent (PGD) and momentum iterative fast gradient sign method (MI-FGSM). For the setting of hyper-parameters, we set the maximum perturbation to be $\epsilon = 16$ in all experiments with pixel values in [0,255]. For the iterative attack methods, we set the maximum number of iteration as 20 and the step size as $\alpha = 1.6$. For MI-FGSM, we set the default decay factor $\mu = 1.0$.

A. IMAGE CLASSIFICATION

In this paper, we use 10 models to make prediction on 3000 images of the CAAD dataset. Table 2 shows the classification accuracy of these models. As we can see from the table, these models can classify the images with high accuracy more than 91.9% and the best one achieves 99.966% accuracy. This implies the normally trained models and the adversarial models can classify the original examples correctly.

B. RESTRICTED REGION ATTACK

In our experiments, we use four methods (Guassian noise, Max-Value-Add, Max-Value-Sub and Average Perturbation) to conduct the restrict region attacks. Figure 4 shows some generated images when we conduct attacks on the restricted region. As we can see, the added perturbations are quite small and we can also recognize them correctly. We compare the prediction accuracy of the ten models on the generated adversarial images. As shown in Table 3, the prediction accuracy is still high because we only apply some easy attack methods, but the accuracy decreases compared to the results in Table 2 where no attacks are performed. From Table 3, the average perturbation method works best among them, which can reduce the accuracy by 3% to 8% compared with other perturbation methods. Therefore, we use the average perturbation method in the restricted region in the following experiments.

TABLE 3. Comparison of four different perturbation methods on the restricted region.

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
Guassian	96.366%	93.233%	92.400%	91.733%	**91.233%**
Max_Value_Add	95.733%	92.666%	91.133%	90.933%	**90.866%**
Max_Value_Sub	95.266%	92.166%	92.000%	91.066%	**90.333%**
Average	90.500%	90.000%	89.333%	88.500%	**87.933%**
Defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
Guassian	94.466%	91.666%	**91.566%**	91.800%	91.833%
Max_Value_Add	96.100%	91.800%	92.300%	**91.266%**	92.033%
Max_Value_Sub	95.633%	**91.233%**	92.166%	91.600%	92.366%
Average	91.766%	89.400%	**88.866%**	89.300%	90.333%

TABLE 4. Different gradient-based (InceptionV3) Attacks on ten models.

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
FGSM	**36.066%**	72.866%	74.400%	72.333%	70.233%
PGD	**0.000%**	86.466%	88.900%	88.700%	87.700%
MI-FGSM	**1.266%**	64.633%	74.500%	74.900%	72.066%
Defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
FGSM	**79.333%**	87.200%	88.066%	80.266%	89.700%
PGD	94.366%	90.866%	91.600%	**90.566%**	91.566%
MI-FGSM	**77.600%**	78.966%	80.933%	78.133%	82.266%

TABLE 5. Attack Comparison of different iterations on the RRI-GM method (FGSM).

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
1 iteration	**32.766%**	66.100%	67.366%	64.966%	61.033%
2 iterations	**12.766%**	46.566%	50.300%	44.900%	42.700%
3 iterations	**5.866%**	30.166%	30.366%	28.833%	26.200%
Defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
1 iteration	**68.133%**	83.166%	83.166%	73.833%	86.233%
2 iterations	**42.566%**	66.133%	67.000%	56.366%	65.466%
3 iterations	**22.200%**	44.666%	46.666%	36.166%	41.966%

C. GRADIENT-BASED ATTACK

We choose three different gradient-based attack methods to conduct our experiments: FGSM, PGD and MI-FGSM. As depicted in Fig. 1, we use InceptionV3 model to calculate the gradient to generate adversarial examples.

By generating the adversarial examples, we can use them to attack the other models and we show the results in Table 4. From the table, the prediction accuracy decreases dramatically by the gradient-based attack method against the InceptionV3 model. However, we can see that the prediction accuracy of some models are still high, which implies the attacks are not that successful; we will show more results by our method.

We depict some examples after we adopt gradient-based attack methods in Fig. 5. The first row of images are the original images and the adversarial images generated by FGSM, the second row of images are generated by PGD method and MI-FGSM respectively. From the figure, the added perturbation is small and we can recognize them easily. However, the neural networks may output incorrect results.

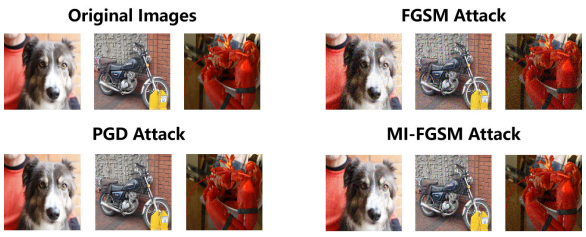


FIGURE 5. Original images and adversarial images by the gradient-based attack methods (FGSM, PGD, MI-FGSM) on the restricted region.

D. RESTRICT REGION BASED ITERATIVE GRADIENT ATTACK

In order to reduce the accuracy of image classification for most black-box models, we propose the restrict region based iterative gradient attack method. We use average perturbation to the restrict region and then we adopt gradient-based attack methods to generate adversarial examples. We conduct these steps for many epochs/iterations. As shown in Table 5, Table 6 and Table 7 when we adopt FGSM, PGD and MI-FGSM respectively, we can conclude that the method can reduce the accuracy of image classification for most black-box models. Among these methods, MI-FGSM iterative attack outperforms others.

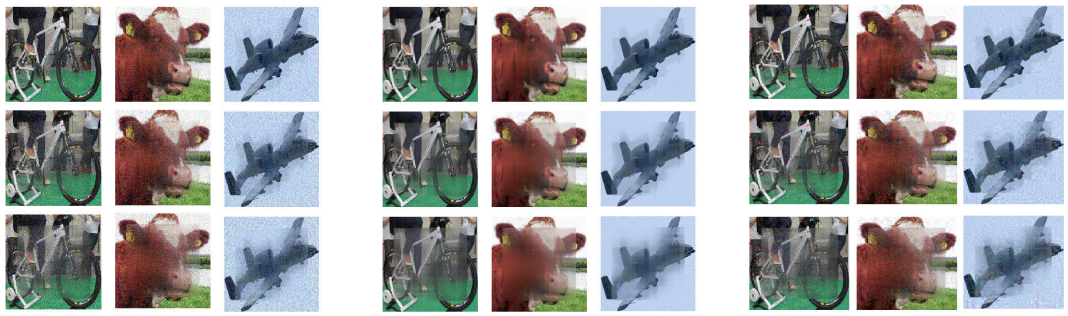
We also show some generated adversarial examples by the restricted region based iterative gradient method in Fig. 6. By adopting FGSM, Fig. 6(a) shows the generated adversarial examples. The first row of images are the adversarial images generated by our method with only one iteration, the second row of images are the adversarial images generated by two iterations, while the third row of images are generated by three iterations. From the comparison, more iterations would add more perturbation but we can still recognize them easily.

TABLE 6. Attack Comparison of different iterations on the RRI-GM method (PGD).

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
1 iteration	**0.000%**	78.666%	83.333%	82.033%	80.333%
2 iterations	**0.003%**	62.866%	66.300%	65.633%	63.800%
3 iterations	**0.000%**	42.233%	44.100%	44.200%	40.733%
defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
1 iteration	89.366%	**86.900%**	88.300%	87.600%	89.266%
2 iterations	**74.600%**	75.566%	76.233%	77.000%	79.566%
3 iterations	**50.566%**	54.633%	56.000%	56.966%	61.033%

TABLE 7. Attack Comparison of different iterations on the RRI-GM method (MI-FGSM).

Normally trained model	InceptionV3	InceptionV4	ResNetV2-152	ResNetV2-101	ResNetV2-50
1 iteration	**0.333%**	53.166%	65.133%	63.300%	60.833%
2 iterations	**0.000%**	28.633%	40.800%	40.466%	36.433%
3 iterations	**0.000%**	14.533%	23.233%	22.966%	21.300%
Defense models	AdvInceptionV3	Ens3AdvInceptionV3	Ens4AdvInceptionV3	AdvInceptionResNetV2	EnsAdvInceptionResNetV2
1 iteration	**67.400%**	71.433%	72.866%	70.266%	74.733%
2 iterations	**48.433%**	49.433%	52.400%	50.733%	55.333%
3 iterations	**28.266%**	28.700%	32.366%	30.266%	34.333%



(a) FGSM (b) PGD (c) MI-FGSM

FIGURE 6. Original images and adversarial images by the RRI-GM (FGSM, PGD, MI-FGSM respectively).

From Table 5, more iterations could generate adversarial examples that have better attack performance even against black-box defense models. For normally trained models, adversarial examples generated by this method can reduce the accuracy to 25% approximately (the best one reduces it to 5.86%). Even for defence models, the accuracy of image classification is reduced to about 38%. The effect is very remarkable for improving the transferability of adversarial examples against black-box models by our method.

We show some generated examples by adopting PGD attack in our method as Fig. 6(b). The examples show similar trend where more perturbation is added to the restricted region with more iterations, but the attack performance becomes much better.As shown in Table 6, adversarial examples generated by this method can reduce the accuracy to 27% approximately for normally trained models. Even for defence models with adversarial learning, the accuracy of image classification is reduced to about 55%.

Similarly, we show the generated examples by adopting MI-FGSM as Fig. 6(c) and Table 7 shows the attack performance. For normally trained models, the generated adversarial examples by our method can reduce the accuracy to 18% approximately and the accuracy of image classification is reduced to about 30% even for defence models with adversarial learning. Among these tables, adopting MI-FGSM could achieve best attack performance. With only three iterations, all these models could be attacked with high success rate.

VI. ADVANTAGES AND DISADVANTAGES

In this paper, we propose the restricted region based iterative gradient attack method (RRI-GM) to generate adversarial examples that have good performance against black-box neural networks. The experimental results show that the generated adversarial examples could attack both normally trained models (without adversarial learning) and defence models with high success rate. This implies that the method could improve the transferability of the adversarial examples. Furthermore, our method is very simple, which only ensembles two methods to generate adversarial examples that fool most neural networks. In addition, our method only adds small perturbation since we restrict the change of each pixel within 16 units.

There also exist some issues to be explored in the future to improve our method. First, the object detection algorithms

cannot always generate the key regions since we have to define the important parts/objects beforehand. Second, although generating the adversarial examples with more iterations could achieve better performance, it incurs more time to realize the attack. Hence it would be interesting and important the explore the tradeoff between the attack performance and the efficiency.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose the restricted region based iterative gradient attack method to generate adversarial examples that have higher transferability against the black-box normally trained models and defense models. We conduct a lot of experiments to validate the effectiveness of proposed method. The best adversarial examples generated by the restricted region based iterative gradient (MI-FGSM) attack can fool all 10 models in our experiment. The results imply the vulnerability of current black-box neural networks and we need to pay more attention to the robustness of neural networks. In the future, we are to improve the key region identification method and explore the tradeoff between the attack performance and the efficiency.

REFERENCES

[1] Z. Tian, C. Luo, J. Qiu, X. Du, and M. Guizani, "A distributed deep learning system for web attack detection on edge devices," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 1963–1971, Mar. 2020, doi: 10.1109/TII.2019.2938778.

[2] Z. Tian, S. Su, W. Shi, X. Du, M. Guizani, and X. Yu, "A data-driven method for future Internet route decision modeling," *Future Gener. Comput. Syst.*, vol. 95, pp. 212–220, Jun. 2019.

[3] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–11.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.

[5] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2017, *arXiv:1611.01236*. [Online]. Available: https://arxiv.org/abs/1611.01236

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[7] S. T. K. Jan, J. Messou, Y.-C. Lin, J.-B. Huang, and G. Wang, "Connecting the digital and physical world: Improving the robustness of adversarial attacks," in *Proc. AAAI*, 2019, pp. 1–8.

[8] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1625–1634.

[9] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," 2017, *arXiv:1707.07397*. [Online]. Available: https://arxiv.org/abs/1707.07397

[10] C. Zhu, V. C. M. Leung, K. Wang, L. T. Yang, and Y. Zhang, "Multi-method data delivery for green sensor-cloud," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 176–182, May 2017.

[11] C. Zhu, L. Shu, V. C. M. Leung, S. Guo, Y. Zhang, and L. T. Yang, "Secure multimedia big data in trust-assisted sensor-cloud for smart city," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 24–30, Dec. 2017.

[12] Z. Tian, W. Shi, Y. Wang, C. Zhu, X. Du, S. Su, Y. Sun, and N. Guizani, "Real-time lateral movement detection based on evidence reasoning network for edge computing environment," *IEEE Trans. Ind. Informat.*, vol. 15, no. 7, pp. 4285–4294, Jul. 2019.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018, pp. 1–28.

[14] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*. [Online]. Available: https://arxiv.org/abs/1605.07277

[15] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2016, *arXiv:1608.04644*. [Online]. Available: https://arxiv.org/abs/1608.04644

[16] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[17] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, 2017, pp. 506–519.

[18] F. Tramer, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," 2017, *arXiv:1704.03453*. [Online]. Available: https://arxiv.org/abs/1704.03453

[19] F. Tramer, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. ICLR*, 2018, pp. 1–20.

[20] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," 2018, *arXiv:1802.00420*. [Online]. Available: https://arxiv.org/abs/1802.00420

[21] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. ICLR*, 2017, pp. 1–24.

[22] N. Papernot, P. Mcdaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 582–597.

[23] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–9.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. CVPR*, 2014, pp. 1–14.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*. [Online]. Available: https://arxiv.org/abs/1512.03385

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.

[27] J. Y. Lee and F. Dernoncourt, "Sequential short-text classification with recurrent and convolutional neural networks," 2016, *arXiv:1603.03827*. [Online]. Available: https://arxiv.org/abs/1603.03827

[28] V. Passricha and R. K. Aggarwal, "Convolutional neural networks for raw speech recognition," in *From Natural to Artificial Intelligence-Algorithms and Applications*. IntechOpen, 2018, doi: 10.5772/intechopen.80026.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–9.

[30] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.

[31] X. Dong, W. Zhang, and N. Yu, "CAAD 2018: Powerful none-access black-box attack based on adversarial transformation network," 2018, *arXiv:1811.01225*. [Online]. Available: https://arxiv.org/abs/1811.01225

[32] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1378–1387.

[33] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 86–94.

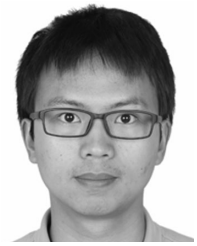
[34] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," 2015, *arXiv:1511.05122*. [Online]. Available: https://arxiv.org/abs/1511.05122

[35] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1–10.

[36] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

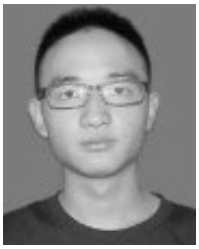
[37] S. Sarkar, A. Bansal, U. Mahbub, and R. Chellappa, "UPSET and ANGRI: Breaking high performance image classifiers," 2017, *arXiv:1707.01159*. [Online]. Available: https://arxiv.org/abs/1707.01159

[38] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, 2017, pp. 1–13.
[39] N. Narodytska and S. P. Kasiviswanathan, "Simple black-box adversarial perturbations for deep networks," 2016, *arXiv:1612.06299*. [Online]. Available: https://arxiv.org/abs/1612.06299
[40] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, and D. Song, "Physical adversarial examples for object detectors," 2018, *arXiv:1807.07769*. [Online]. Available: https://arxiv.org/abs/1807.07769
[41] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1528–1540.
[42] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Adversarial generative nets: Neural network attacks on state-of-the-art face recognition," 2017, *arXiv:1801.00349*. [Online]. Available: https://arxiv.org/abs/1801.00349
[43] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace Face ID system," 2019, *arXiv:1908.08705*. [Online]. Available: https://arxiv.org/abs/1908.08705
[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 1–7.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.



ZHAOQUAN GU received the bachelor's and Ph.D. degrees in computer science from Tsinghua University, in 2011 and 2015, respectively. He is currently a Professor with the Cyberspace Institute of Advanced Technology (CIAT), Guangzhou University, China. His research interests include wireless networks, distributed computing, big data analysis, and artificial intelligence security.

WEIXIONG HU received the bachelor's degree in computer science and technology from NanChang University, China, in 2018. He is currently pursuing the master's degree in computer science and technology with Guangzhou University, China. His research interests span deep learning, image classification, and adversarial examples.



CHUANJING ZHANG received the bachelor's degree in computer science and technology from Northeastern University, China, in 2018. He is currently pursuing the master's degree in computer science and technology with Guangzhou University, China. His research interests span deep learning, image classification, and adversarial examples.

LE WANG received the Ph.D. degree in computer science from NUDT. He is currently an Associate Professor with the Cyberspace Institute of Advanced Technology, Guangzhou University. His current research interests include networks and big data security. He was a member of the China Computer Federation.

CHUNSHENG ZHU received the Ph.D. degree in electrical and computer engineering from The University of British Columbia, Canada. He is currently an Associate Professor with the SUSTech Institute of Future Networks, Southern University of Science and Technology, China. He is also an Associate Researcher with the Peng Cheng Laboratory, PCL Research Center of Networks and Communications, China. He has authored more than 100 publications published by refereed international journals (e.g., the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, IEEE TRANSACTIONS ON COMPUTERS, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING, IEEE TRANSACTIONS ON CLOUD COMPUTING, *ACM Transactions on Embedded Computing Systems*, and *ACM Transactions on Cyber-Physical Systems*), magazines (e.g., the *IEEE Communications Magazine, IEEE Wireless Communications Magazine*, and *IEEE Network Magazine*), and conferences (e.g., the IEEE INFOCOM, IEEE IECON, IEEE SECON, IEEE DCOSS, IEEE ICC, and IEEE GLOBECOM). His research interests mainly include the Internet of Things, wireless sensor networks, cloud computing, big data, social networks, and security.

ZHIHONG TIAN is currently a Professor, a Ph.D. Supervisor, and the Dean of the Cyberspace Institute of Advanced Technology, Guangzhou University. He is also the standing Director of CyberSecurity Association of China. From 2003 to 2016, he worked at the Harbin Institute of Technology. His current research interests are computer networks and network security. He is a member of China Computer Federation.